2014-03

IPv6 geolocation using latency constraints

Tran, Tony V.H.

Monterey, California: Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

IPv6 GEOLOCATION USING LATENCY CONSTRAINTS

by

Tony V. H. Tran

March 2014

Thesis Advisor:	Robert Beverly
Second Reader:	Geoffrey G. Xie

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704–0188*

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704–0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202–4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD–MM–YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From — To)*
27–03–2014	Master's Thesis	03-06-2012 to 03-28-2014

4. TITLE AND SUBTITLE

IPv6 GEOLOCATION USING LATENCY CONSTRAINTS

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Tony V. H. Tran

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Postgraduate School
Monterey, CA 93943

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Department of the Navy

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

The views expressed in this document are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.IRB Protocol Number:N/A.

14. ABSTRACT

IPv4 addresses are now exhausted, and as a result, the growth of IPv6 addresses has increased significantly since 2010.The rate of increase of IPv6 usage is expected to continue; thus the need to determine the geographic location of IPv6 hosts will grow to support location-aware applications.Examples of services that require or benefit from IPv6 geolocation include overlay networks, location-based security mechanisms, client language and policy determination, and location targeted advertising.Internet protocol (IP) geolocation is the process of obtaining the geographical location of a device or host using only the host's IP address.This study looked at using constraint-based geolocation (CBG), a latency-based measurement technique, on IPv6 infrastructure and analyzed location accuracy against ground truth.Results show that overall IPv6 CBG had up to 30% larger average error distance estimates as compared to IPv4 CBG. However, CBG performance varied depending on the location of the target host.Hosts located in the Asia-Pacific region performed the worst, while hosts located in Europe had the best performance in median error distance.AS-level path differences between IPv4 and IPv6 and the number of landmarks had the most significant impact on CBG performance.

15. SUBJECT TERMS

IPv6, Geolocation, Multilateration, Delay measurements

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			
Unclassified	Unclassified	Unclassified	UU	65	**19b. TELEPHONE NUMBER** *(include area code)*

THIS PAGE INTENTIONALLY LEFT BLANK

IPv6 GEOLOCATION USING LATENCY CONSTRAINTS

Tony V. H. Tran
Lieutenant, U.S. NAVY
B.S., University of Oklahoma, 2005

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

from the

NAVAL POSTGRADUATE SCHOOL
March 2014

Author:	Tony V. H. Tran
Approved by:	Robert Beverly Thesis Advisor
	Geoffrey G. Xie Second Reader
	Peter J. Denning Chair, Department of Computer Science

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

IPv4 addresses are now exhausted, and as a result, the growth of IPv6 addresses has increased significantly since 2010. The rate of increase of IPv6 usage is expected to continue; thus the need to determine the geographic location of IPv6 hosts will grow to support location-aware applications. Examples of services that require or benefit from IPv6 geolocation include overlay networks, location-based security mechanisms, client language and policy determination, and location targeted advertising. Internet protocol (IP) geolocation is the process of obtaining the geographical location of a device or host using only the host's IP address. This study looked at using constraint-based geolocation (CBG), a latency-based measurement technique, on IPv6 infrastructure and analyzed location accuracy against ground truth. Results show that overall IPv6 CBG had up to 30% larger average error distance estimates as compared to IPv4 CBG. However, CBG performance varied depending on the location of the target host. Hosts located in the Asia-Pacific region performed the worst, while hosts located in Europe had the best performance in median error distance. AS-level path differences between IPv4 and IPv6 and the number of landmarks had the most significant impact on CBG performance.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

Ark	archipelago
AS	autonomous system
ASN	autonomous system number
BGP	Border Gateway Protocol
CAIDA	Cooperative Association for Internet Data Analysis
CDF	cumulative distribution function
CIDR	Classless Inter-domain Routing
CBG	Constraint Based Geolocation
DNS	Domain Name System
EDD	error distance delta
GPS	Global Positioning System
ICMP	Internet Control Message Protocol
IP	Internet Protocol
IPv4	Internet Protocol version 4
IPv6	Internet Protocol version 6
IANA	Internet Assigned Numbers Authority
ISP	Internet Service Provider
NAT	Network Address Translation
PCC	Pearson Correlation Coefficient
ping	Packet InterNet Groper
PL	path length
RIR	Regional Internet Registry
RR	resource record
RTT	round trip time
TCP	Transmission Control Protocol

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgements

I want to thank Professor Robert Beverly for guiding me through all aspects of this thesis. There were times when I was not sure how I was going to accomplish something, and he always seemed to have an answer or a direction to guide my wandering efforts. Thank you for this opportunity to work with you. Without your guidance and insight, this work would not have been possible.

I also like to thank Professor Geoffrey Xie for providing insightful feedback and making sure I made this thesis easier to understand. The ability to leverage his expertise was beneficial throughout this research.

Lastly, I want to thank my family for their unwavering support throughout my life. You have taught me many valuable lessons that no academic institution could teach. I appreciate your constant support, consideration, and love.

Always in my thoughts, however far away I am. WE DID IT!

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1: Introduction

An Internet Protocol (IP) address is a unique 32 or 128 bit unsigned integer assigned to every device connected to the Internet. The Internet Assigned Numbers Authority (IANA) is the governing body that allocates IP address blocks to the Regional Internet Registrys (RIRs). Different RIRs serve the following areas of the world: Africa, Asia/Pacific, North America, Latin America, Europe/Middle East/Central Asia. The five RIRs then further allocate IP address blocks to Internet Service Providers (ISPs) that are then assigned to businesses, organizations and individuals. Similar to a postal address, an IP address is required to properly request and receive data between Internet devices. 32-bit Internet Protocol version 4 (IPv4) addresses provide roughly 4 billion addresses; however, because of address allocation policy that depends on contiguous address blocks for route aggregation, the number of available IPv4 addresses is much smaller. In 2011, IANA's pool of available IPv4 addresses was exhausted, and the RIRs have few addresses remaining [1, 2]. Although Network Address Translation (NAT) [3], which permits multiple devices to communicate using a single public IP address, has historically relieved some of the IPv4 address pressure, it has long-term limitations. NAT requires at least one public address, a requirement that is become more difficult to meet in large residential networks and in countries without large address allocations. Further, NAT impedes end-to-end connectivity, creates single-points of failure where fates are shared, and implies potential collisions in large enterprises connecting VPNs [4].

Now with the exhaustion of unassigned IPv4 addresses, the industry is more rapidly moving toward IPv6, including adoption by large infrastructure operators and network providers [1, 2, 5, 6]. Internet Protocol version 6 (IPv6) was standardized in 1998 as the successor to IPv4, the Internet's long-standing Internet protocol [7]. Major network vendors and operators already support IPv6 in their operating systems and network equipment [6, 8]. Additionally, the U.S. government has mandated that their networks shift to IPv6 to increase network robustness and mission capability [8].

As IPv6 grows, so does the need to determine the geographic location of these connected hosts. IP geolocation is the process of obtaining the physical geographical location of a device or host on the basis of its IP address. The geographic granularity can span continents, countries, regions, cities, or streets [9]. Location-aware applications depend on efficient and accurate inference of IP geolocation to drive a multitude of valuable services. Services requiring geolocation include content delivery providers that rely on the location of their users; transaction authorization from approved locations; and automated selection of browser/content features like language and consumer advertising. A world where every connected devices could be located would enable countless innovative services.

1.1 Motivation

The growth of IPv6 continues to require research into network infrastructure, topology, and supporting services. Internet users that do not have IPv6 dual-stack support, and thus cannot reach IPv6 sites directly must use IPv4 infrastructure to carry IPv6 packets. This is done using a technique known as tunneling, which encapsulates IPv6 packets within IPv4. Peering agreements, Domain Name System (DNS) services, traffic and workload types in IPv6 can also vary compared to IPv4 [10]. Unlike NAT in IPv4, where multiple devices can take on a single IP address, the address space for IPv6 is so massive it is likely that a single device will take on multiple IPv6 addresses. A study conducted by Dhamdhere et al. [11] suggests that data plane performance of IPv6 is comparable to that of IPv4 if AS-level paths are the same, but can be much worse than IPv4 if the AS-level paths differ. To the best of our knowledge, how this difference affects current delay-based methods of IPv6 geolocation has not been investigated. These differences mentioned above could reveal that some methods for geolocating IPv4 may not work for Today, much of IP geolocation research has been focused on IPv4. Previous work in geolocating IPv6 hosts were largely unfulfilled due to low IPv6 host density and lack of supporting infrastructure [12]. Thus, we explore IPv6 geolocation in today's Internet using the Constraint Based Geolocation (CBG) methodology, a latency-based approach.

1.2 Research Questions

This thesis explores whether using latency-based measurements is a viable "first-step" coarse-grain geolocation technique for IPv6 hosts. We create inference models, which help us estimate distance to a target, using multilateration of known hosts to geolocate our collected ground truth datasets. In doing so, we investigate the following:

- What is the accuracy of CBG when geolocating IPv6 hosts?
- What are the accuracy differences for dual-stacked hosts?
- What other unknown factors could affect the accuracy of CBG when geolocating IPv6 hosts?

1.3 Thesis Structure

The remainder of this thesis is organized as follows:

- Chapter 2 covers leading IP geolocation techniques in use and related work.
- Chapter 3 discusses the CBG technique, multilateration process and how delay measurements convert to distance constraints.
- Chapter 4 details the results from geolocating and measuring the AS-level paths of three datasets. One dataset was manually collected by confirming IPv4-v6 address pairs of academic universities or institutions. Further details on how we collected this dataset is discussed in Chapter 4. The second dataset is a listing of dual-stacked servers with known location provided by a content distribution network. Our last dataset comes from [13], listing one-to-one IPv4-v6 address pairs. However, true location is unknown for these hosts. We indirectly measure accuracy by comparing the estimated CBG IPv4 to the estimated CBG IPv6 location, which provides a measure of confidence to our CBG estimates.
- Chapter 5 provides conclusions based on this research and recommendations for future areas of research.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 2:
Background and Related Work

There are many ways to obtain geographical data from an Internet host. An obvious way would be to ask the device or end-user to submit data of where they are located and provide updates if their location changes or one could utilize the Global Positioning System (GPS) on that device to provide locational data. But what if the device or end-user does not want to share their geographic location data due to privacy concerns or provides false data? Or what if the device is not equipped with a GPS sensor or is not activated? Moreover, infrastructure devices such as routers, switches, or servers may also need to be geolocated. Clearly, obtaining accurate and reliable locational data is difficult and poses several challenges. Ideal IP geolocation methods strive to determine the location of an end device with the least effort and resources, while generating the most accurate result. A process that can be automated and continually updated is also ideal. This chapter reviews the leading methods in obtaining geographical data on Internet hosts and discusses the challenges each method presents.

2.1 Geolocation Methods

There are two broad categories to IP geolocation. The first is a database-driven approach containing records for a range of IP addresses. Geographical information is associated with each range of IP addresses. Some IP geolocation databases are fee for use, and others can be searched for free online. The second category relies on network measurements, where delays and information of the network topology is used. In this study, we look at CBG, a delay-based approach.

2.1.1 Commercial and Public Sources

IP geolocation databases contain records for a range of IP addresses called *blocks* or *prefixes*. Blocks can span non-Classless Inter-domain Routing (CIDR) subsets of the address. Most geolocation database entries are composed of an integer value pair corresponding to an address block range. Each block is associated with geographical information that could include country, region, state, city, ZIP code, area code, and latitude/longitude. Users can then query the database to obtain geographical data regarding an IP address [14]. Some examples of IP geolocation databases are IP2Location, MaxMind, HostIP, and InfoDB [14, 15]. IP2Location and MaxMind are paid service commercial databases, while HostIP and InfoDB are freely available. Methods into how commercial databases perform IP geolocation are proprietary so little is known of how these work. InfoDB is built upon a free MaxMind database version and is incremented by the IANA locality information. Lastly, HostIP data relies on direct feedback from participating users and ISPs [14, 16]. Table 2.1 is an overview of the numbers of records within each geographic database [17].

Database	Address Blocks	Lat Long	Countries	Cities
HostIP	8,892,291	33,680	238	23,700
IP2Location	6,709,973	17,183	240	13,690
InfoDB	3,539,029	169,209	237	98,143
MaxMind	3,562,204	203,255	244	175,035

Table 2.1: Number of entries recorded within each geographic database

A study conducted by Uhlig et al. show the geolocation accuracy for InfoDB and MaxMind have roughly the same distance distributions since InfoDB is based on the free version of MaxMind [14]. About 20% of the blocks in both InfoDB and MaxMind have error distances under 20 km from the ground truth. The remaining 80% have between 20 km and 800 km, where 800 km was considered the maximum distance in a country and cut off at that distance. IP2Location had larger error distances with roughly 20% of the blocks under 200km and the remaining 80% between 200 km and 800 km. Another study into IP geolocation databases conducted by Huffaker et al. show the HostIP database performing with varying results depending on the IP targets. PlanetLab, a globally distributed set of computers available for computer networking research and distributed systems research, were used and showed about 79% of addresses within 80 km of ground truth [18]. A dataset from Freebox that lists French ADSL networks by region showed HostIP database with 80% of addresses over 100 km, with 20% over 1000 km [18].

Using public information sources and commercial databases has its limitations. Geolocation databases may make us of public information sources such as *whois* or DNS. The whois service is a Transmission Control Protocol (TCP)-based transaction-oriented query and response protocol used to provide information services to internet users. This service is like a “white-pages” for registered domains, allowing users to request registered information by the organization such as the organization’s phone number, administrator, or physical address [19]. The DNS is a mechanism for naming resources where names are usable in different hosts, network and administrative organizations. A DNS resource record (RR) provides a mapping of hostname to an IP address [20]. Public databases are not completely reliable since there are no requirements or incentives to keep it up-to-date and accurate. A single organization such as an ISP can be allocated an IP block, but will only register one location such as their corporate headquarters.

This effectively maps large address blocks to a single location. Of course, not all end hosts are located at or near the organization's registered address. This can introduce error distances to the end host if the ISP has consumers in geographically dispersed areas. Commercial databases also leverage public sources and ISP block allocations to populate their databases but can fall to the same problem of stale or bad data entries. Without knowing the methodology of commercial databases in performing IP geolocation, their reliability also comes into question. Studies into geolocation databases show that these databases have made significant improvements and can usually geolocate at the country-level. However, database entries are greatly disproportionate to popular countries (e.g., U.S.) and are not viable for consistent and accurate geolocation services [14, 21].

2.1.2 Measurements

The advantage of a latency-based approach is that it is coupled to a device's location by physics, whereas the database method is not. Delay measurements are the most widely used and improved upon among geolocation measurement techniques. Probe tools, such as Packet InterNet Groper (ping) or traceroute, are network administration tools used to test the reachability of a host on an IP network and to measure the round trip time (RTT) between an originating and destination host. An Internet Control Message Protocol (ICMP) echo request packet is sent from say a source host A and waits for the corresponding ICMP echo reply packet from the destination host B [22]. The time delta of the probes is the RTT delay measurements. Figure 2.1 show the process.

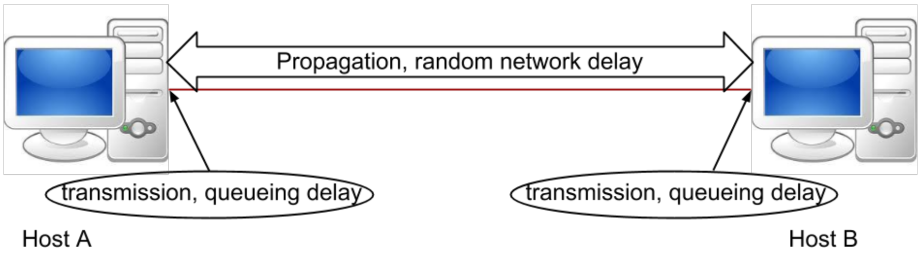
Propagation, random network delay
transmission, queueing delay
transmission, queueing delay
Host A
Host B

Figure 2.1: Overall delay between two random hosts in packet-switched network

The overall delay, d, between two randomly selected hosts in a packet-switched network is expressed below:

$$d = d_t + d_p + q + \varepsilon \tag{2.1}$$

Transmission delay d_t is the time the first and last bit leaving the outbound link on one end of a host. The propagation delay d_p is the physical minimum time the bits take traveling on the wire to reach the input link of the receiving host. Queuing delay, q, is the time a packet is queued for transmission, but waits for the output link to be available. Random delay, ε, is time loss due to media access contention, router processing overhead, and network disturbances. We assume that propagation delay dominates transmission delay and processing delay. Packets traveling along a fiber-optic cable are estimated to move at two-thirds the speed of light in a vacuum [23]. This provides a convenient value of 1ms RTT per 100km of cable to express the geographic distance in light-milliseconds to obtain an absolute maximum distance using the RTT (or one-way delay) between A and B. The measured RTT is always larger than the aforementioned ”perfect-case” because fiber-optic cable is not laid out in a straight line. Cable paths usually follow railways, highways, or power lines, through varying levels of elevation producing a rather circuitous path through any number of nodes. This is far from the straight-line ideal case and thus introduces delay in the from of added distance [23]. Lastly, BGP inter-AS routing policies tend to exhibit path inflation generating larger delays resulting in larger geographic distances. Therefore, a geolocation technique should account for network topology and routes to capture path-specific latency inflations [12].

The GeoPing method developed by Padmanabhan and Subramanian [24] uses network delay measurements made from geographically dispersed locations to infer the coordinates of the end host or nearby neighbors. GeoPing does this by building a map of delay vectors from a set of probes with known location we will call landmarks, to a single host also with known location. A delay vector to an end host, or target, with unknown location is measured from the set of all probes to the target. The delay vector is compared to every delay vector in the map of delay vectors to identify the approximate match. The Euclidean distance between the delay vector and every other vector is calculated, and the “nearest” landmark to the distance vector is selected as the location estimate of the target. The intuition is that the delay that packets experience between network hosts is a function of the geographic distance between the hosts. Landmarks are known positions of end devices that are measured to determine a network delay model. The measured delay pattern of the target host is compared to the collected landmark models and an location estimation for the target host is produced. GeoPing is not without limitations. Possible location estimates for a target is dependent on the number of landmarks participating. Thus, this limits the location estimate to a discrete set of possible locations [25, 26].

The CBG methodology builds off the work of GeoPing and introduces multilateration to infer the location of Internet hosts. The CBG methodology will be detailed in Section 3.

2.2 Geolocation in IPv6

2.2.1 Parallels to IPv4

IPv6 can use the same public and commercial databases used in IPv4. We could associate an IPv6 address to an IPv4 address through whois or DNS, which can provide more locational data to the IPv6 address. Also, as we will later find in Chapter 4, commercial organizations appear to be leveraging existing IPv4 address entries, to provide geographical location data to associated IPv6 addresses.

IPv6 still travels over the same medium as IPv4 and is subject to the same physical constraints. Delay over IPv6 paths is comparable to that over IPv4 paths if AS-level paths are the same [11]. As a starting point for measurement-based IPv6 geolocation, we use the same latency-based techniques used for IPv4 geolocation.

2.2.2 Past Work

The only prior study that used delay measurements to geolocate IPv6 devices was conducted in 2006 [12]. Only two measurement nodes in western Europe that supported IPv6 were available for testing. Delay measurements for IPv6 were measured between these two nodes and compared to IPv4 delay between the same two nodes. Then, an adjustment factor was applied to IPv4 measurements to create an artificial data set for IPv6 CBG. Their results did not explicitly state error distance performance of IPv6 against IPv4, but only that the IPv6 delay factor over IPv4 was 1.06 suggesting that the confidence area regions would increase by the same factor. The study conducted by [12] compared IPv6 performance against IPv4 was summarized as still incomplete due to the lack of IPv6 supporting infrastructure and landmarks.

2.2.3 Challenges

Geolocating in the IPv6 poses different and potentially harder challenges over IPv4. Unlike IPv4, IPv6 has a much larger address space, making it infeasible to enumerate and record geolocation data for all potential IPv6 address blocks. Also, the affects from tunneling and auto-tunneling through transition technologies such as 6to4, Teredo, or NAT64/DNS64, are still unknown for delay-based geolocation methods [10]. Lastly, although data plane performances of IPv6 are comparable to that of IPv4 if AS-level paths are the same, this does not confirm that we can use delay-based geolocation for the IPv6 space [11]. For AS-level paths that do differ, the feasibility and level of accuracy when using delay-based geolocation for IPv6 is still unknown.

We know that the number of globally unique IPv6 ASs is now roughly 17% compared to IPv4 [27]. Recent work has shown that the average AS path length for IPv6 is decreasing. The work concludes that the overall decreasing trend on the average IPv6 AS path length is due to an increasing dominance of a single internet transit provide [11]. Thesis thesis looks at how IPv4 and IPv6 path lengths (PLs) affect CBG.

CHAPTER 3:
Methodology

IP geolocation is the process of obtaining the physical geographic location of a device given only the device's IP address. CBG is a popular first-step method for coarse-grained IPv4 geolocation and serves as a foundation for other, more precise, geolocation methods, e.g., [9, 12, 26, 28, 29]. CBG infers the geographic location of end hosts using multilateration. Multilateration refers to the process of estimating a position using multiple distance measurements from known landmarks to the target. For example, GPS multilateration requires at least three satellites to estimate the position of a GPS receiver. Precise timing kept by satellites with on-board atomic clocks provide the receiver timing and timing interval information to calculate its position. Each satellite continually transmits messages to the GPS receiver that includes the time the message was transmitted, and the satellite's position at time of message transmission. The receiver determines the transit time of each message and computes the distance to each satellite using the speed of light. These distances and satellites' locations are used to compute the location of the receiver using analytic geometry [30].

Unlike GeoPing, described in Section 2.1.2, where the possible inferences of a device's geolocation are limited to a discrete set of locations, CBG estimates the target location within a constrained continuous space. Additionally, a confidence region to the estimated location of the target is provided, allowing for geographic area resolution analysis. Lastly, CBG allows for the relationship between network delay and distance to be *re-calibrated* given the current state of the network. This is achieved through regular network delay measurements to calculate the geographic distance relationship.

This chapter first describes our method of categorizing landmarks and hosts by regions. Next, the network infrastructure from which we obtain network delay measurements is discussed. Then, we describe the CBG methodology to geolocate an arbitrary target IP. Finally, we detail how we used AS-level path measurements to provide insight into CBG performance.

3.1 Geographical Layout

We categorized each landmark and host from our datasets described in Section 4.2 using known ground truth or by inferred MaxMind location by country and by geographic region for comparison. Our categorization is the same used by the RIR [31] and is shown in Figure 3.1 with one exception. We separate an area from the APNIC region into a region we will refer to as Oceania. We outline Oceania as the southeast corner of Asia to include Australia and land areas not connected to mainland Asia. We note that all hosts located in Malaysia are located on the Asian mainland, not on Malaysia's island territory. In separating the APNIC region into two regions, we hope to provide more insight into the Asia Pacific islands that are geographically not connected to the Asian mainland. This geographical feature could reveal significant differences in latency measurements, thus affecting CBG performance in each region. In total, there are six regions and 84 countries among all datasets.

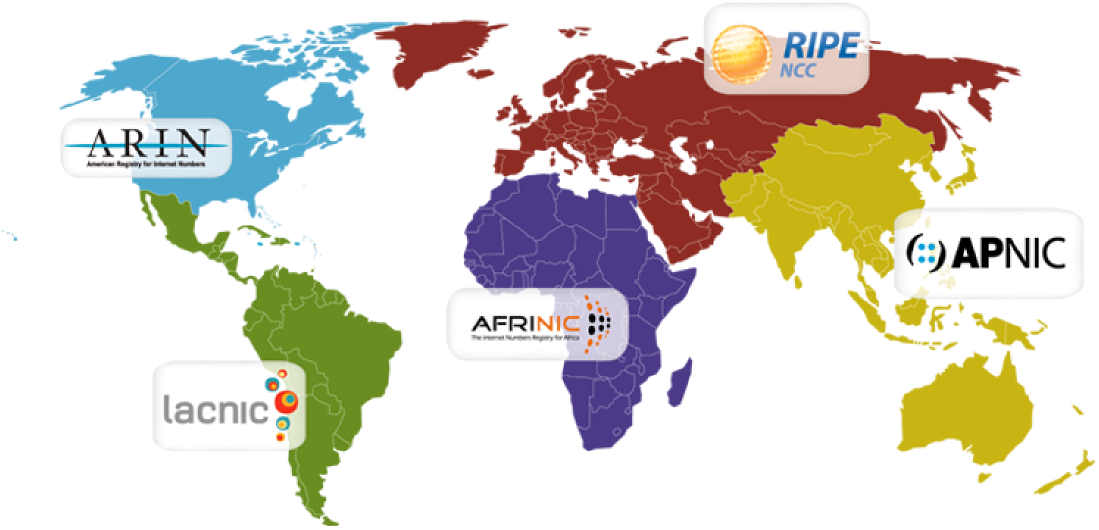
RIPE NCC
ARIN
American Registry for Internet Numbers
APNIC
AFRINIC
The Internet Numbers Registry for Africa
lacnic

Figure 3.1: Global map of Regional Internet Registry coverage

The countries summarized in Table 3.1 represent all the countries found in our dataset; however, not all countries were represented in each dataset. Table 3.1 serves as an aid to the reader to how we categorized a landmark or host, particularly in Oceania, the Caribbean, or bordering regions. Additionally, it serves as a minimum list of countries represented in this study.

AfriNIC	***APNIC***	***RIPE NCC***	***LACNIC***	***ARIN***	***Oceania***
Egypt Kenya Mauritius South Africa Tanzania	Bhutan China Hong Kong India Japan Malaysia Singapore South Korea Sri Lanka Taiwan Thailand Vietnam	Austria Bahrain Belgium Bosnia/Herzegovina Bulgaria Croatia Cyprus Czech Republic Denmark Estonia Finland France Germany Greece Hungary Iceland Ireland Israel Italy Kazakhstan Kuwait Latvia Lichtenstein Luxembourg Macedonia Moldova Netherlands Norway Oman Poland Portugal Qatar Romania Russian Serbia Slovakia Slovenia Spain Sweden Switzerland Turkey Ukraine United Arab Emirates United Kingdom Vatican City	Argentina Brazil Chile Colombia Costa Rica Ecuador Honduras Mexico Peru Sint Maarten Uruguay	Bahamas Canada Grenada Jamaica Puerto Rico United States	Australia Indonesia New Caledonia New Zealand Philippines

Table 3.1: Country to region category dataset

3.2 Probing Infrastructure

For our experiment, we utilize the worldwide distributed network measurement infrastructure of Cooperative Association for Internet Data Analysis (CAIDA). CAIDA is a collaboration among commercial, government, and research organizations to promote greater cooperation in the engineering and maintenance of global Internet infrastructure. archipelago (Ark) is CAIDA's active measurement platform for scientific analysis of Internet traffic, topology, routing, and performance [32]. The geographic location of the Ark monitors are known and serve as our landmarks from which we build our network delay models. Of the 80 monitors within the Ark infrastructure, 29 monitors were both IPv4 and IPv6 capable. Figure 3.2 shows the locations of the Ark monitors used in this study.

Due to our constrained list of capable IPv4-v6 landmarks, this study is limited to a set of landmarks predominantly located in the United States and Western Europe. As such, we expect that without landmarks in Central & South America, Africa, and the Middle East regions, this will degrade our IP geolocation performance for hosts located in those regions. Table 3.2 shows a complete listing and details of the Ark landmarks used in this study.

As described in Section 2.1.2, probes will be sent from our select group of landmarks to our target hosts in each of our datasets described in Chapter 4. To measure RTT, ping probes will be utilized. For measuring AS-level paths, traceroute probes will be used. Section 3.5 describes how we utilized traceroute probes to measure AS-level paths.

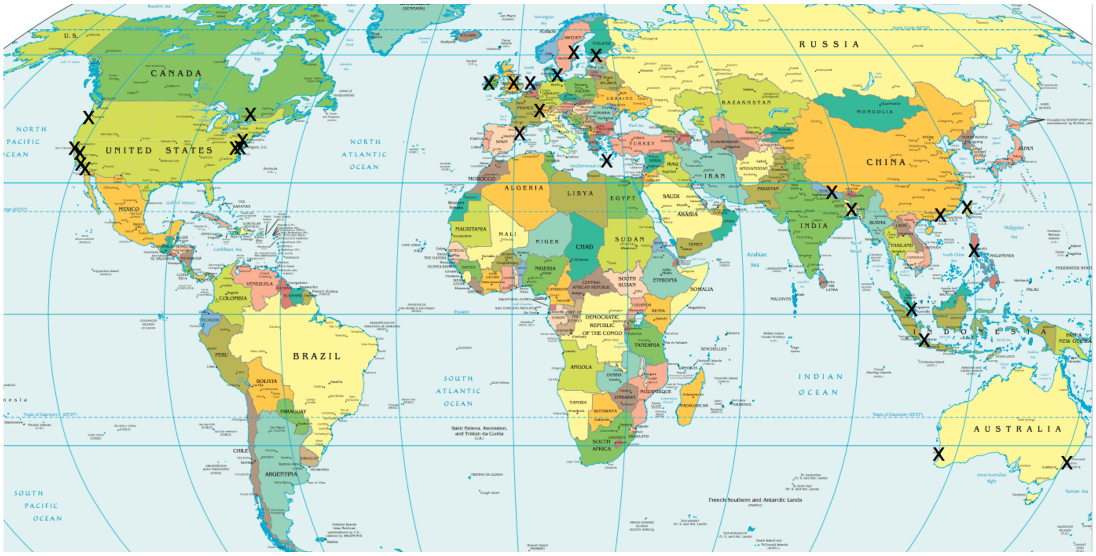

Figure 3.2: Location of Archipelago landmarks

Landmark Name	Location	Organization	Region
ams-nl	Amsterdam, Netherlands	SURFnet	RIPE NCC
ams2-nl	Amsterdam, Netherlands	AMS-IX	RIPE NCC
ams3-nl	Amsterdam, Netherlands	RIPE NCC	RIPE NCC
bcn-es	Barcelona, Spain	Universitat Politecnica de Catalunya	RIPE NCC
bma-se	Kista, Sweden	Acreo	RIPE NCC
bwi-us	Aberdeen, MD, U.S.	U.S. Army Research Lab	ARIN
cbg-uk	Cambridge, United Kingdom	University of Cambridge	RIPE NCC
cgk-id	Jakarta, Indonesia	Indonesian IPv6 Task Force	Oceania
cph-dk	Ballerup, Denmark	Solido Networks ApS	RIPE NCC
dac-bd	Dhaka, Bangladesh	BDCOM Online Limited	APNIC
dub-ie	Dublin, Ireland	HEAnet	RIPE NCC
eug-us	Eugene, OR, U.S.	University of Oregon	ARIN
hel-fi	Espoo, Finland	TKK	RIPE NCC
her-gr	Heraklion, Crete, Greece	Foundation for Research and Technology	RIPE NCC
hkg-cn	Hong Kong, China	Tinet	APNIC
iad-us	Chantilly, VA, U.S.	ARIN	ARIN
jfk-us	New York, NY, U.S.	Hurricane Electric	ARIN
ktm-np	Kathmandu, Nepal	Nepal Research and Education Network	APNIC
lax-us	Los Angeles, CA, U.S.	CENIC	ARIN
mnl-ph	Quezon City, Philippines	Advanced Science Technology Institute	Oceania
per-au	Perth, Australia	AARNet	Oceania
san-us	San Diego, CA, U.S.	CAIDA	ARIN
sin-sg	Singapore, Singapore	DCS1 Pte Ltd	APNIC
sjc2-us	San Jose, CA, U.S.	Hurricane Electric	ARIN
sql-us	Redwood City, CA, U.S.	Internet Systems Consortium	ARIN
syd-au	Sydney, Australia	AARNet	Oceania
tpe-tw	Hsinchu, Taiwan	TWAREN	APNIC
yow-ca	Ottawa, ON, Canada	Ottawa Internet Exchange	ARIN
zrh2-ch	Zug, Switzerland	Kantonsschule Zug	RIPE NCC

Table 3.2: Archipelago landmarks used to probe in both IPv4/6 space

3.3 Constraint-based Geolocation

3.3.1 Building Landmark Delay Models

Landmark delay models establish great-circles, or range circles, that are used to convert delay measurements to geographic distance constraints [25]. This is achieved by establishing an maximum, or loose distance, bound called the *baseline* and tight distance bound called the *bestline* for each landmark L_i in our set of landmarks L_N as shown in Table 3.2. This relationship is described in slope-intercept form by $y = mx + b$ where m is the speed that the bits travel on the medium and b denotes the delay factor experienced on the network. The estimated geographic distance measured in kilometers (km) is represented by x. The RTT measured in milliseconds (ms) is represented by y. The baseline and bestline is further explained below.

The baseline is the theoretical "perfect case" and applies to each of our landmarks. As described in Section 2.1.2, we use 2/3 the speed of light as the speed digital information travels along a fiber-optic cable in a vacuum, which translates nicely to $m_{base} = 1/100$. Since the baseline is our perfect case with no additive delay, we set $b_{base} = 0$. CBG deliberately makes a distance overestimation to the target for the baseline to ensure that the solution space is not empty. Using known m_{base} and b_{base}, we compute the loose bound distance between sites by simply solving for distance x from our baseline model with the minimum RTT measured between the two sites. As shown from Figure 3.3, the baseline greatly overestimates the distance to a target host.

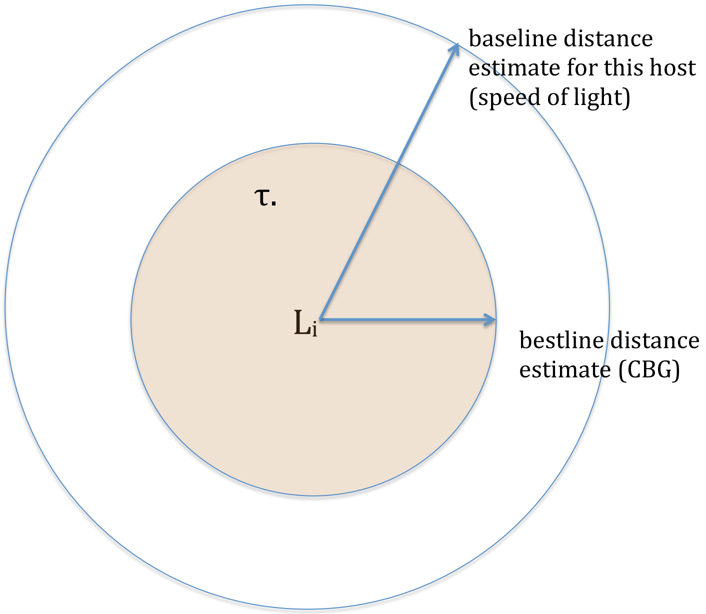
baseline distance estimate for this host (speed of light)
τ.
Li
bestline distance estimate (CBG)

Figure 3.3: Baseline and bestline greater circle target distance estimation

To tighten this range and try to account for the additive network delay, we borrowed the work from [25] to compute a bestline $y = m_i x + b_i$ for each landmark L_i. The input consists of known pairwise geodistances between landmarks, denoted by g_{ij}, and minimum measured pairwise RTTs between landmarks, denoted by d_{ij}. Figure 3.3 shows range constraint circles using baseline and bestline models for a target host τ from L_i.

To find m_i and b_i we use an objective function that minimizes the distance between the bestline with non-negative y-intercept against all delay measurements from a given landmark L_i. This function is expressed as:

$$minimize \sum_{\forall j \neq i} (d_{ij} - m_i g_{ij} - b_i) \tag{3.1}$$

subject to $b_i \geq 0$; $m_i \geq m_{base}$; $d_{ij} \geq m_i g_{ij} + b_i$, $\forall j \neq i$. With Equation 3.1, we solve for the two unknowns m_i and b_i for the landmark L_i. Specifically, we select the smallest RTT delay experienced from that landmark L_i and solve for the unknowns at this delay value, and all subsequent delays with distances greater than d_{ij}. This becomes a linear programming problem. Figure 3.4 shows an example IPv4 delay distance scatter-plot with the loose bound "baseline" and tight bound "bestline" for a landmark located in New York City, NY. The resulting bestline $y = m_i + b_i$ is the line closest to, but below all the measured distance-delay measurement points x, y that meet the provided constraints that the b_i is non-negative with the smallest m_i. Section 3.3.3 explains how we use landmark bestline models to geolocate target hosts.

In cases where b_i is negative, b_i is set to zero and evaluated to see if the bestline is below all the delay-distance measurement points x, y. If it is, then b_i is set to zero with its corresponding m_i. If the bestline is not below all the distance-delay measurement points x, y, the bestline is set to baseline values.

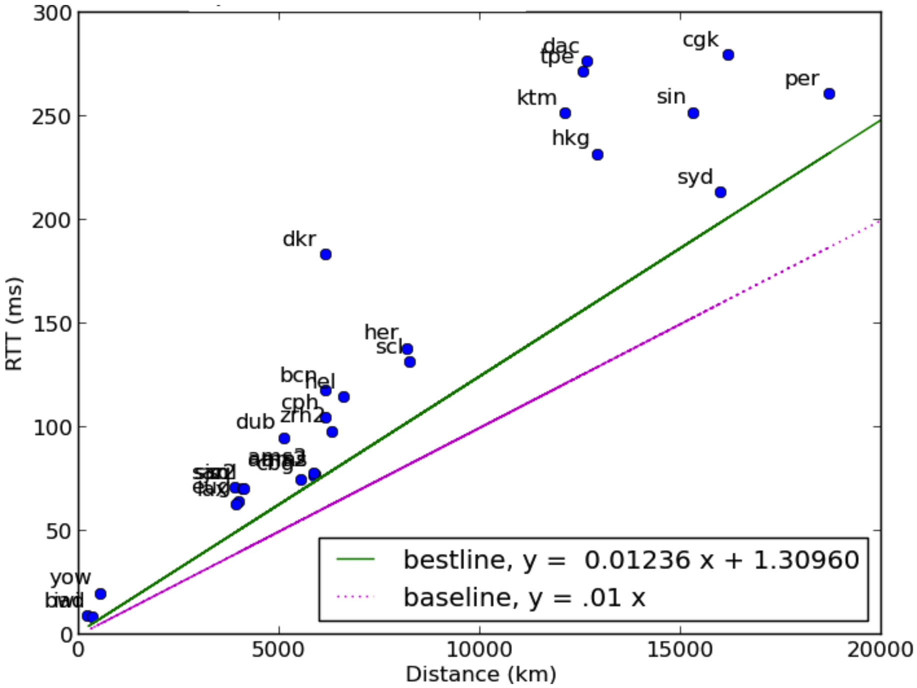

Figure 3.4: Landmark model of geographic distance and network delay

3.3.2 Building Ark Landmark Delay Models

To generate the bestline for each Ark landmark L_i, we used the administrative network utility ping described in Section 2.1.2 to measure network delay. Four measuring sessions were conducted over a four-month period at different segments of a day where RTTs were recorded between an Ark landmark L_i and all other Ark landmarks L_N using ping probes. A measuring session consisted of nine ping probes sent from an Ark landmark L_i to all other Ark landmarks L_N. This step was repeated for all Ark landmarks L_N in our dataset. In total, up to 36 RTTs were recorded between any two Ark landmarks. After all RTT measurements were collected over the four sessions, the smallest RTT measurement observed between an Ark landmark L_i and Ark landmark L_j was recorded to build the bestline model. This effectively captures the overall minimum network delay between two landmarks during this 4 month period, which helps provide the tightest bound bestline for each of our landmark.

This procedure was conducted on our Ark landmarks probing both their IPv4 and IPv6 addresses.

In total, 29 IPv4 and 29 IPv6 bestline models for each landmark from Table 3.2 were generated, corresponding to our 29 Ark landmarks. We note that as the state of the network is dynamic, each landmark bestline model can be *re-calibrated* through re-probing of the network delay and geographic distance. In doing this, the overall accuracy of CBG improves and reflects the most current state of the network. In Chapter 4, we provide our bestline model results for each of our Ark landmarks.

3.3.3 Using Great Circle Constraints to Geolocate Hosts

The CBG methodology uses multilateration with geographic distance constraints based on delay measurements to infer the location of Internet hosts. The estimated geographic distance constraint $\hat{g}_{i\tau}$ between a landmark L_i and target τ is derived from the measured delay $d_{i\tau}$ using the bestline model of the landmark L_i. In other words, each landmark L_i uses its own bestline equation to calculate a $\hat{g}_{i\tau}$ to a target τ using $d_{i\tau}$. This is expressed in Equation 3.2.

$$\hat{g}_{i\tau} = \frac{d_{i\tau} - b_i}{m_i} \tag{3.2}$$

This constraint represents a great-circle $C_{i\tau}$ with the landmark L_i at its center and $\hat{g}_{i\tau}$ the radius. Figure 3.5 shows CBG with three landmarks. Each landmark has no sense of direction to where the target is located, only that the target is within the circumference of its circle and the estimated distance to it. Particularly, the estimated distance is always an overestimation since there is always an additive distortion inherent in network delay measurements as described in Section 2.1.2.

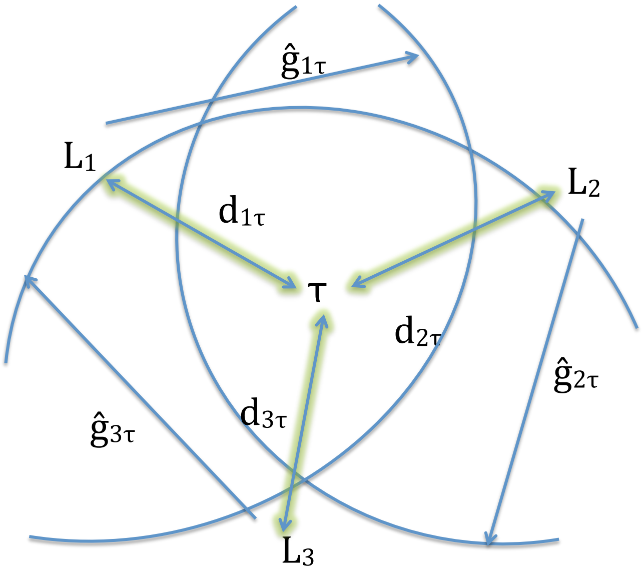

Figure 3.5: Basic CBG using three landmarks

The intersection of all the circles $C_{i\tau}$ to a target τ generates an area region R where the target is believed to be in and is defined in Equation 3.3, where K is the total number of landmarks. This can be seen as an order-K Venn diagram, where a given a set of landmarks K produce a collection of circles $C_{i\tau}$ to geolocate a target τ.

$$R = \bigcap_{i=1}^{K} C_{i\tau} \tag{3.3}$$

To define the physical boundaries of R, which is a convex hull, all the intersection points from the great-circle constraints $C_{i\tau}$'s for a given target τ is collected. The set of points that fall out of the most restrictive, or smallest circle, are eliminated. The surviving intersecting points are then sorted and used to find a series of line segments that intersect the circle to isolate only the points for our desired polygon region to estimate the target location. The area of a R on a sphere is calculated using Equation 3.4 [33]. Each intersecting point v is a pair of latitude ϕ_v and longitude λ_v points.

$$A = -\frac{R^2}{2} \sum_{v=0}^{N} (\lambda_{v+1} - \lambda_{v-1}) \cdot sin\phi_i \tag{3.4}$$

Per the CBG method, the centroid of the region R is chosen as the target's estimated location. To calculate the centroid coordinates (c_x, c_y) of the target τ in R, we perform an approximation using Equation 3.5 and Equation 3.6 where | **M** | denotes the determinant of the matrix.

$$c_x = \frac{1}{6A} \sum_{n=0}^{N-1} (x_n + x_{n+1}) \begin{vmatrix} x_n & x_{n+1} \\ y_n & y_{n+1} \end{vmatrix} \tag{3.5}$$

$$c_y = \frac{1}{6A} \sum_{n=0}^{N-1} (y_n + y_{n+1}) \begin{vmatrix} x_n & x_{n+1} \\ y_n & y_{n+1} \end{vmatrix} \tag{3.6}$$

Figure 3.6 is an example of great-circle distance constraints $C_{i\tau}$'s geolocating a host τ over Western Europe. The shaded orange region is the intersection region R.

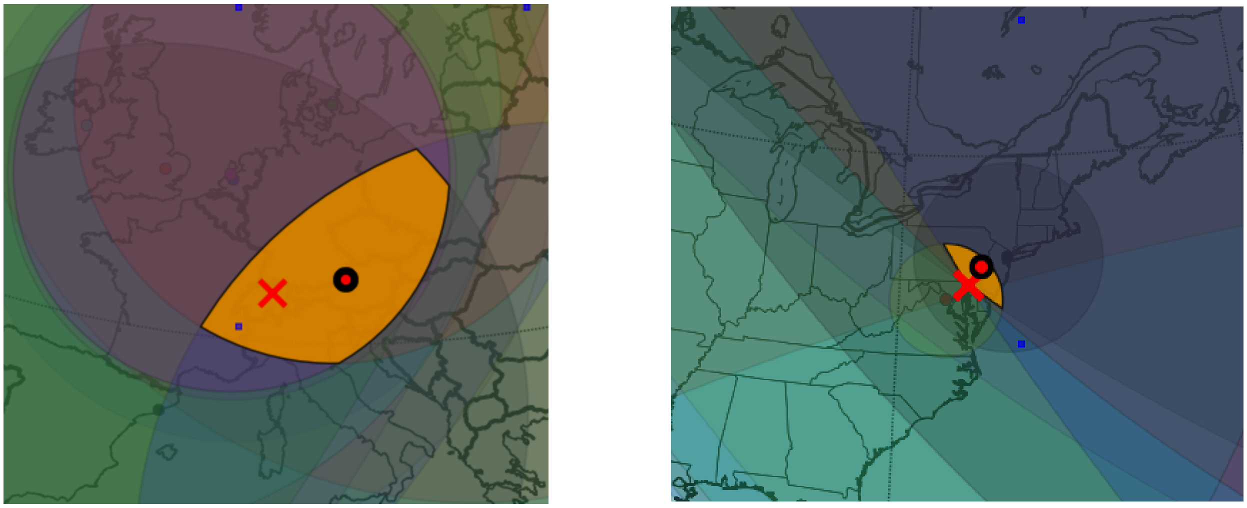

Figure 3.6: Location estimation of target

After inferring the point estimate for each target, the centroid coordinates, we compute the error distance, which is the distance difference between the centroid estimate and the actual target τ location. In Chapter 4, we show our results comparing IPv4 and IPv6 CBG geolocation among three datasets.

3.4 Confidence Regions

The intersected region R provides a confidence level to the estimated position of the target τ. Intuitively, the area of R quantifies the geographic extent of each host location estimate in km^2. When comparing the two regions in Figure 3.6, we can see that the smaller the area, the more confident our estimation to the target [25]. We use confidence regions to provide location estimation resolution and another quantitative metric to compare IPv4/6 geolocation performance. In Chapter 4 we discuss our CBG geolocation results for our collected datasets.

3.5 AS-level Path Measurements

Studies have shown that IPv4 and IPv6 path similarity, or congruence, is correlated with performance [11]. To explore the trends in congruity for AS-level paths, we look at the edit distances and path lengths of each IPv4 and IPv6 forward AS-level path. First, to find the forward AS-level path, all IP address hops between from each landmark to a target host is recorded from a traceroute probe. Next, RouteViews Border Gateway Protocol (BGP) tables are used to match IPv4/6 addresses to the corresponding autonomous system number (ASN). RouteViews is a project founded by Advanced Network Technology Center at the University of Oregon to allow

Internet users to view global BGP routing information from the perspective of other locations around the internet. RouteViews servers receive their information by peering directly with other BGP routers, typically at large internet exchange points [34]. Since IPv4/6 address hops recorded in the traceroute may belong to the same AS, duplicate ASN are not recorded in the forward AS-level path between a monitor and target host. Thus, only unique ASes on that path are used to compute edit distance and path length. Table 3.3 is an example of a AS-level path with ASN values between a monitor that probed a single target in IPv4 and IPv6. We note that inferring the AS path from IP router interfaces is not perfectly accurate [35], but suffices for the purposes of our large-scale analysis that seeks to understand the AS path to CBG relationship.

3.5.1 Comparing Paths Using Edit Distances

To compare IPv4 and IPv6 AS-level paths, we use the same technique by Dhamdhere et al. to compute the number of AS changes, or edits, required to make the IPv4 path identical to the IPv6 path [11]. Identical paths will have a value of zero edits since there are no changes needed to the ASs in the IPv4 path to make it the same as the IPv6 AS path. The higher the edit distance, the more the paths differ. For example, Table 3.3 has an edit distance of three. We note that additions to a path "shift" the AS-level path to the right. So for the IPv4 AS-level path to match the IPv6 AS-level path, the following edits would be made: add AS 11537 at position A, change AS 668 to AS 13 at position C, then lastly add AS 6022 at position D.

Hop	A	B	C	D	E
IPv4 AS	5050	668	3999		
IPv6 AS	11537	5050	13	6022	3999

Table 3.3: Example AS-level path comparison

3.5.2 Comparing Paths Using Path Lengths

The path length of an AS-level path is the number of unique ASs in the AS-level path. From Table 3.3, the IPv4 AS-level path has a length of three and the IPv6 AS-level path has a length of five. We will compute the average PL by region and country to determine any correlation this has with CBG performance.

CHAPTER 4:
Experimental Results

The goal of this thesis was to explore IPv6 CBG performance compared to CBG IPv4. IPv4-v6 address pairs were collected to provide a base for comparison. In doing so, we developed bestline delay models for our Ark landmarks to describe the current state network delay experienced from each landmark. Then, we conducted RTT measurements from our landmarks towards hosts within our datasets. Using the Ark landmark bestline models we developed and collected RTT data from landmark to host, we geolocated each host using the CBG method described in Chapter 3. We also investigated how AS-level path differences between IPv4 and IPv6, and AS-level path lengths affect IP CBG performance.

In the results below, we found among our datasets that overall IPv6 CBG geolocated targets with error distances and area regions larger than IPv4 CBG, but some regions had much greater error distances than others. The worst case was the Oceania region where CBG IPv6 median error distance was double that of CBG IPv4 median error distance. The best case was the RIPE region where CBG IPv6 median error distance outperformed CBG IPv4 median error distance by roughly 8%.

Using AS-level paths between landmarks and hosts, we used edit distances and PLs to provide insight into our geolocation performance. We found in each dataset that edit distances seemed to have a direct relationship with CBG accuracy. The greater the edit distances, the greater the error distance to the target hosts. We also found that PL did not show a strong relationship with CBG accuracy.

4.1 Ark Bestline Models

Using the procedures described in Section 3.3, we created 29 IPv4 and 29 IPv6 bestline models for each Ark landmark representing the relationship between the current network delay and geographic distance were generated. We also calculated the Pearson Correlation Coefficient (PCC) of each model. In statistics, the PCC is a measure of the linear correlation (dependence) between two variables X and Y, giving a value between +1 and −1 inclusive, where 1 is total positive correlation, zero is no correlation, and −1 is total negative correlation. It is widely used in the sciences as a measure of the degree of linear dependence between two variables [36, 37]. For our landmark bestline models, the PCC measures the dependence between RTT and geographic distance. Figure 4.1 shows the baseline and bestline model of an IPv4 Ark landmark located in Aberdeen, MD, U.S. with a high PCC of 0.9360. Figure 4.2 shows the baseline and bestline model of an IPv6 Ark landmark located in Hong Kong, China, with a negative PCC (inverse relationship) of −0.7518.

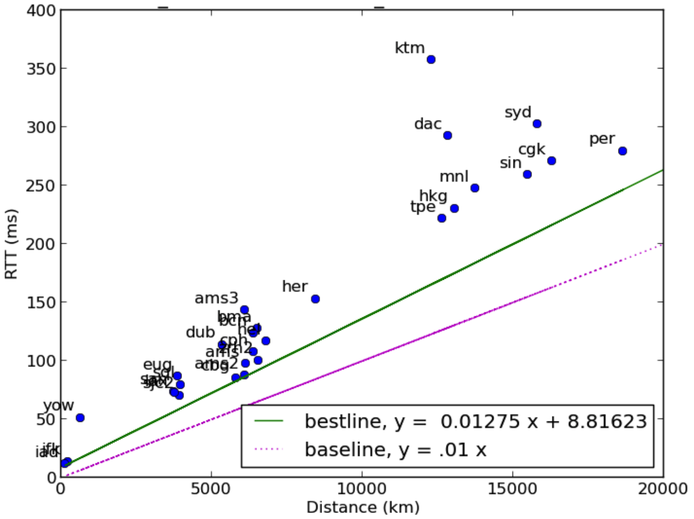

Figure 4.1: Landmark *bwi* with high PCC

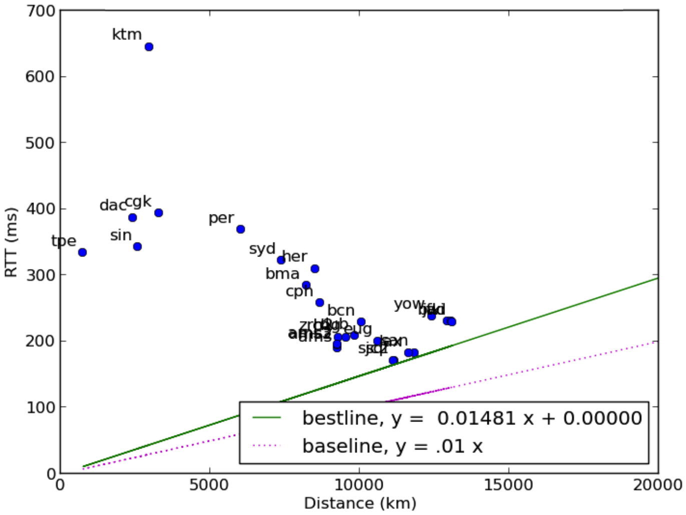

Figure 4.2: Landmark *hkg* with low PCC

Table 4.1 shows our bestline results for IPv4 and IPv6 along with the PCC. We point out that all negative PCC for both IPv4 and IPv6 are landmarks located in the APNIC and Oceania regions. This indicates a poor correlation and dependence between RTT and geographic distance, thus making it more difficult to accurately geolocate hosts from these landmarks.

	IPv4			***IPv6***		
Landmark	m_i	b_i	PCC	m_i	b_i	PCC
ams	0.0113	0.3413	0.9333	0.0113	0.2433	0.8848
ams2	0.0113	0.3521	0.8951	0.0114	0.2706	0.9298
ams3	0.0100	0.0000	0.8794	0.0100	0.0000	0.9297
bcn	0.0135	11.3463	0.9055	0.0138	9.1904	0.8558
bma	0.0145	6.8992	0.8705	0.0152	7.4310	0.9017
bwi	0.0128	8.816	0.8875	0.0119	5.3649	0.9360
cbg	0.0123	2.6073	0.9077	0.0108	3.0331	0.8810
cgk	0.0120	4.2376	0.2866	0.0100	24.6548	0.1529
cph	0.0142	1.8542	0.9177	0.0132	2.3132	0.6112
dac	0.0161	6.5948	0.3349	0.0135	28.7467	-0.2614
dub	0.0150	1.8687	0.9101	0.0150	1.9122	0.8867
eug	0.0103	5.6024	0.8459	0.0100	0.0000	0.9095
hel	0.0100	0.0000	0.8591	0.0100	0.0000	0.7733
her	0.0100	0.0000	0.7757	0.0100	0.0000	0.7449
hkg	0.0137	0.0000	-0.2397	0.0148	0.0000	-0.7518
iad	0.0123	1.6550	0.9522	0.0114	1.9994	0.7020
jfk	0.0100	0.0000	0.9473	0.0108	2.1358	0.7062
ktm	0.0199	0.0000	0.2346	0.0147	47.9024	-0.0104
lax	0.0100	0.0000	0.9121	0.0130	1.3225	0.9196
mnl	0.0140	3.0674	0.5465	0.0100	0.0000	-0.1875
per	0.0118	2.8490	0.0831	0.0128	0.8105	-0.1225
san	0.0100	6.0577	0.9046	0.0108	1.5634	0.9277
sin	0.0110	4.0304	0.2642	0.0100	24.5755	-0.1424
sjc2	0.0120	0.3828	0.8872	0.0120	0.4080	0.6804
sql	0.0120	0.3798	0.9028	0.0120	0.3791	0.7722
syd	0.0103	5.8406	0.4668	0.0105	5.3352	0.3739
tpe	0.0100	0.0000	0.2764	0.0100	0.0000	0.1520
yow	0.0123	6.7287	0.8937	0.0125	4.4403	0.8987
zrh2	0.0100	0.0000	0.8700	0.0119	3.3081	0.8479

Table 4.1: Archipelago landmark bestline models

4.2 Target Host Datasets

This study uses three different datasets of IPv4 and IPv6 IP addresses as summarized in Table 4.2.. These datasets enable us to independently geolocate the host using its IPv4 and IPv6 addresses for comparison.

For each of our datasets, we conducted probing sessions similar to those described in Section 3.3.2 where each host, playing the role of a "target" to be geolocated, was sent nine ping probes to its IPv4 address and nine ping probes to its IPv6 address. An additional traceroute was used for debugging and to record the AS-level paths. Four probing sessions were conducted over a one month period at different segments of a day and the shortest RTT from a given landmark to a target was then used for geolocation. Using the steps described in Section 3.3.3, we estimate two area regions and two centroids for each target, one based the IPv4 address and one based on the IPv6 address. For our UNI and CDN datasets with known ground truth, we calculate the error distance to each target from the IPv4 and IPv6 centroid coordinates to the actual location coordinates. The error distance delta (EDD) was also calculated which is the absolute error difference between the error distances of IPv4 and IPv6. The EDD allows us to compare how close each estimated IP version geolocation was relative to the other. For our ONE2ONE dataset with unknown ground truth, we calculated the distance between the IPv4 and IPv6 centroid coordinates.

Dataset	No. of Hosts	No. of Regions Represented	No. of Countries Represented	Ground truth known	Description
UNI	53	4	8	Yes	Manually collected
CDN	1940	6	69	Yes	Provided
ONE2ONE	1697	6	69	No	Provided

Table 4.2: Dataset characteristics

For each dataset, we compared our CBG results against the IP geolocation database MaxMind. MaxMind distributes free IPv4 and IPv6 geolocation databases, enabling us to retrieve their estimated location data for an IP address [15]. We note that although MaxMind's free version is advertised to be less accurate than their licensed IP geolocation software, we compared MaxMind against our CBG results as a general comparison of performance.

Lastly, we examined and compared IPv4 and IPv6 AS-level paths to targets to uncover any relationship between AS-level path patterns to CBG performance. Determining the AS path from each vantage point to each target requires running traceroutes and then mapping interface addresses to ASs. We note that both of these processes have errors; some paths force communicating peers to timeout for IPv6 due to routing behavior or ICMP6 filtering. The ability to reach both IPv4 and IPv6 can be as low as 66% [38]. For this study, we limit our analysis on traceroutes that respond to both IPv4-v6 address pairs.

4.3 Academic Institutions (UNI) Dataset

In obtaining our UNI dataset, we relied on the insight that established academic institutions often host their own network infrastructure on-site, especially in the U.S. We first looked at the Allesedv database which lists academic universities and institutions that are IPv6-enabled [39]. Then, we conducted manual DNS queries for AAAA resource records from this list of IPv6-enabled websites. The AAAA record maps an hostname to an IPv6 address similar to how an A record maps an hostname to an IPv4 address. The answer section of the AAAA contains the query answer, the IPv6 address for a given hostname. We discard any hosts whose AAAA records had any indications a third-party was hosting the site (e.g., content distribution provider, web-hosting service). We repeated these steps to obtain the IPv4 address from our list of hostnames. Our target hosts are only those with both IPv4 and IPv6 addresses registered in the DNS. After determining the IPv4 and IPv6 addresses of these institutions, we use the publicly-known location of each site as the actual location of the hosted site. Each physical address is then looked up for its latitudinal and longitudinal coordinates using a web-based geographical coordinate conversion tool [40]. Table 4.3 shows the breakdown of the locations of our UNI target hosts by region and country. The UNI dataset was collected in May 2013.

4.3.1 UNI Geolocation

Table 4.3 shows our IP CBG performance for our UNI dataset alongside results from MaxMind. The EDD is also computed, where EDD is the absolute distance difference between IPv4 and IPv6 error distances to the actual target location. In other words, this is the difference in IPv4-v6 inferred CBG error distance. The EDD provides a perspective on how far IPv6 CBG estimated the host location relative to IPv4 CBG.

Overall, IPv6 CBG average error distance performs roughly 33% worse than CBG IPv4 eon the UNI targets. Yet, the overall IPv6 CBG error distance median is only 17% worse than the IPv4 CBG error distance median. This large difference between average and median is weighted heavily on the much larger error distance experienced in the Oceania region, specifically New Zealand. The RIPE and ARIN regions show comparable CBG performance between IPv4 and

IPv6. RIPE IPv6 CBG average error distance performs around 20% worst for this region against IPv4, yet the median error distance is nearly the same. With only six target hosts, the higher average is due to some hosts with large error distances.

Surprisingly, for the sole country in the APNIC region for this dataset, China, IPv6 CBG geolocation performs roughly 37% better than IPv4 CBG geolocation. When comparing MaxMind's result for this target, MaxMind has zero EDD between IPv4 and IPv6. Taking a closer look, we queried the geographic coordinates for the IPv4-v6 pairs in MaxMind and found coordinates for both IPv4-v6 at the center of the country. It appears that MaxMind in this case assigned the same geographic point for both IPv4 and IPv6. We were unable to discern if MaxMind mapped the IPv6 address to the IPv4 address and returned the location of the IPv4 address, or whether this behavior is a simply a characteristic of the free MaxMind database.

		CBG						***MaxMind***					
Location	No. of Hosts	IPv4 Avg	IPv4 Median	IPv6 Avg	IPv6 Median	EDD Avg	EDD Median	IPv4 Avg	IPv4 Median	IPv6 Avg	IPv6 Median	EDD Avg	EDD Median
APNIC	1	2381	2381	1494	1494	887	887	1143	1143	1143	1143	0	0
RIPE NCC	6	250	248	311	251	73	16	6	6	1488	394	1482	389
ARIN	45	600	484	761	601	326	136	54	5	1528	1745	1474	1587
Oceania	1	2815	2815	13627	13627	10812	10812	0	0	514	514	513	513
China	1	2381	2381	1494	1494	887	887	1143	1143	1143	1143	0	0
Germany	1	358	358	336	336	22	22	10	10	306	306	296	296
Italy	1	384	384	746	746	362	362	3	3	7172	7172	7170	7170
New Zealand	1	2815	2815	13627	13627	10812	10812	0	0	514	514	513	513
Spain	1	22	22	13	13	9	9	5	5	538	538	534	534
Switzerland	1	466	466	462	462	4	4	9	9	122	122	113	113
United Kingdom	2	135	135	156	156	20	20	5	5	394	394	389	389
United States	45	600	484	761	601	326	136	54	5	1528	1745	1474	1587
Overall	53	636	464	967	561	506	136	68	5	1497	1577	1429	1402

Table 4.3: UNI error distances (km) for inferred CBG and MaxMind

Figures 4.3 and Figure 4.4 shows our cumulative distribution function (CDF) error distance and EDD for the UNI dataset. Roughly 80% of both IPv4 and IPv6 CBG geolocated the true target location under 1,000km. MaxMind geolocated to near 80% of targets to 10km or less, which is consistent with their advertised city-level fidelity for IPv4. MaxMind advertises country-level accuracy for IPv6 geolocation which helps explain why IPv6 performs much worse than IPv4.

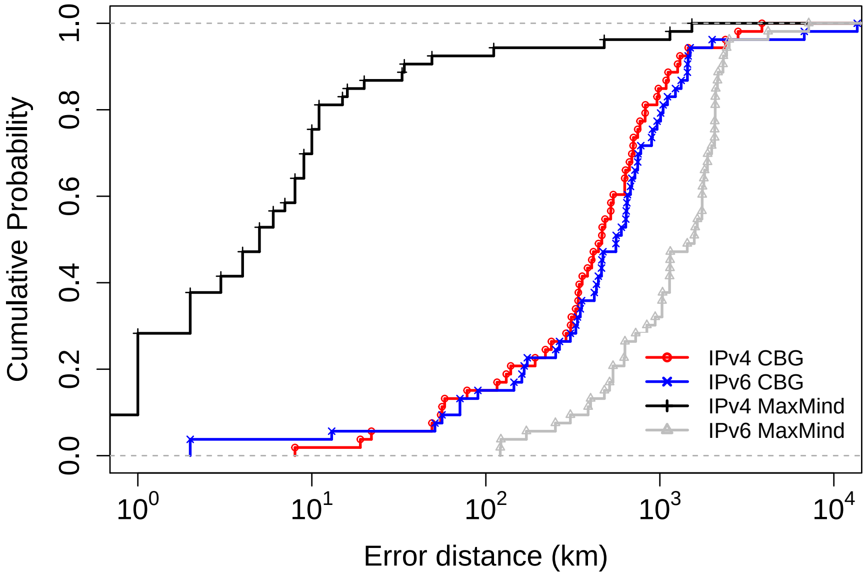

Figure 4.3: CDF error distance for UNI

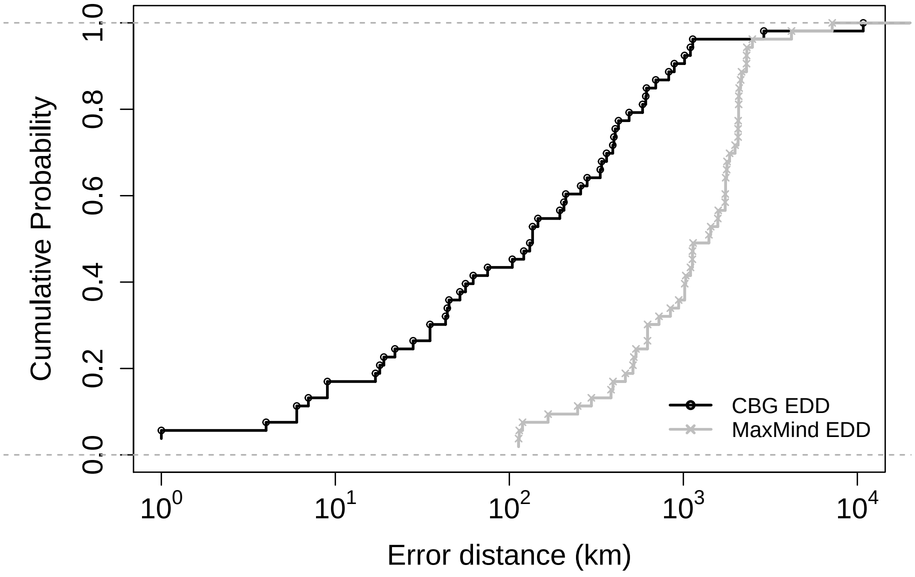

Figure 4.4: CDF error distance difference for UNI

Figure 4.5 shows the confidence regions for IPv4 and IPv6 CBG. We described in Section 3.4 that confidence regions provide an estimated area in which the target is located. The smaller the area, the more confident is the target estimation [25]. We see here that IPv4 areas are slightly smaller than IPv6, providing a slightly higher confidence region. A confidence regions of 10^5 km^2 is roughly the size of the Republic of South Korea or the U.S. state of Kentucky.

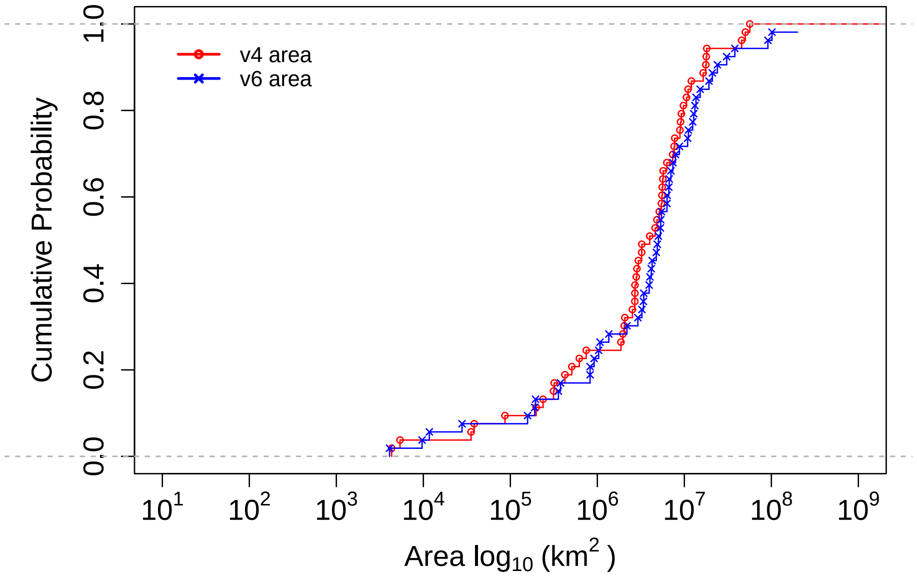

Figure 4.5: CDF Confidence region for UNI CBG

For this dataset, we see overall IPv6 CBG geolocation perform worse than IPv4 CBG geolocation in average error distance and median error distance. Although most regions have few hosts, this dataset still indicates some regions have much better IPv6 CBG geolocation than other regions. It also indicates potentially how well IP CBG performs in certain regions overall.

4.3.2 UNI AS-level Path

Described in Section 3.5, each landmark will send an IPv4 and IPv6 traceroute probe to each host in a dataset. The traceroute probes we collected during our probing sessions are now used to measure the AS-level path. We discard duplicate traceroutes between a landmark and host, resulting in our final IPv4 and IPv6 traceroutes to perform our analysis. In total, we collected 1,484 IPv4 and 1,583 IPv6 unique traceroutes from our landmarks. Of those traceroutes, 6% IPv4 and 14% IPv6 of traceroutes did not get replies from the destination target hosts. We matched 1,360 traceroutes between IPv4 and IPv6 for direct comparison of AS-level edit distance and PL.

Table 4.4 shows the number of traces per country and region. The average edit distance and PL is highest for China, which also has the second largest CBG error distance found in Table 4.3.

New Zealand is not shown below due to traces to that target host not responding. We also find that the overall PL is shorter for IPv6 than IPv4 which is consistent to what was found in recent work discussed in Section 3.5. Note that the average PL for Germany is larger than the United States, but the CBG EDD for Germany is much smaller compared to the United States. The edit distance for Germany is smaller compared to the United States. This leads us to believe that edit distance has a greater affect on CBG performance than PL.

Location	No. of Traces	Edit Distance Average	Edit Distance Median	IPv4 PL Average	IPv6 PL Median
APNIC	27	5	5	5.556	5
RIPE NCC	189	2.085	2	4.143	4
ARIN	1144	2.601	3	4.493	4
China	27	5	5	5.556	5
Germany	27	2.37	3	4.889	5
Italy	27	3.259	3	4.074	4
Spain	54	1.981	2	4.37	4
Switzerland	27	1.741	1	3.889	4
United Kingdom	54	1.63	1	3.704	4
United States	1144	2.601	3	4.493	4
Overall	1360	2.577	3	4.465441	4

Table 4.4: UNI edit distances and path lengths

4.4 Content Distribution Network (CDN) Dataset

Our second dataset is the IPv4 and IPv6 addresses of 1,940 dual-stacked servers from a content distribution network dispersed throughout the globe with known geographic coordinates. The CDN dataset is different from the UNI dataset in that CDN has nearly 40 times as many hosts, and are geographically placed throughout the world. Table 4.5 shows the regions where these hosts are located along with CBG and MaxMind geolocation performance.

4.4.1 CDN Geolocation

Table 4.5 shows our comparison of geolocating IPv6 against the IPv4 pair. The average error distance for IPv4 and IPv6 CBG are 913km and 1380 km, respectively. The median error distance for IPv4 and IPv6 CBG are 516 km and 668 km respectively. Again we find that, overall, IPv6 average error distance performed 33% worst than IPv4, and IPv6 median error distance

23% worse over IPv4. The RIPE and ARIN regions exhibited the best CBG performance. Interestingly, only RIPE IPv6 CBG geolocation median error distance was lower than its IPv4 median error distance. This performance difference may be due to the popularity of IPv6 in the RIPE region, especially in Western Europe. We note that three of the top countries with the largest CBG EDD are from the Oceania region.

AfriNIC and LACNIC performed by far the worse with IPv4 and IPv6 CBG error distance medians three to four times larger than the RIPE NCC error distance medians. Poor performance in these regions is not surprising since most of our landmarks are located in the RIPE and ARIN regions. When comparing AfriNIC CBG performance alone, although the error distances for both IPv4 and IPv6 were well over 1,000km, we find that IPv6 had 39% better geolocation estimation than IPv4. MaxMind performance is worst in the Oceania region. Overall, we find that MaxMind outperforms only CBG IPv4 and IPv6 error distance medians.

		CBG						***MaxMind***					
Location	No. of Hosts	IPv4 Avg	IPv4 Median	IPv6 Avg	IPv6 Median	EDD Avg	EDD Median	IPv4 Avg	IPv4 Median	IPv6 Avg	IPv6 Median	EDD Avg	EDD Median
AfriNIC	30	3795	3345	2328	2435	2016	1844	242	2	1426	509	1415	509
APNIC	186	1230	886	2682	1433	1861	712	401	166	2503	397	2731	517
ARIN	823	571	439	769	583	384	186	1469	874	1418	1074	2185	1838
LACNIC	102	2484	2181	2447	2210	1823	1795	335	130	1552	1216	1567	1114
RIPE NCC	647	667	345	745	319	244	82	505	79	828	247	514	333
Oceania	152	1810	758	4902	1518	3301	724	2003	161	3623	1467	3321	1709
Overall	1940	914	516	1380	668	808	199	1008	231	1505	481	1724	944

Table 4.5: CDN regional error distances (km) for inferred CBG and MaxMind

Table 4.6 lists the top 10 countries in this dataset by largest CBG EDD. We see four countries, United Arab Emirates, Qatar, Bahrain, and Oman are among the top 10 countries located in the RIPE NCC region with large CBG EDD. Also, the IPv4 CBG average error distance for these countries are also large. These four countries are the only four located on the Saudi Arabian peninsula in this dataset.

		Inferred CBG						***MaxMind***					
Location	No. of Hosts	IPv4 Avg	IPv4 Median	IPv6 Avg	IPv6 Median	EDD Avg	EDD Median	IPv4 Avg	IPv4 Median	IPv6 Avg	IPv6 Median	EDD Avg	EDD Median
New Caledonia	3	1932	1932	11973	12291	10041	10360	0	0	130	130	130	130
New Zealand	24	4286	1673	10282	11768	6274	5802	148	105	790	464	895	574
United Arab Emirates	6	2943	3006	7214	7213	4270	4298	36	36	4835	4835	4800	4800
Qatar	3	2717	2827	5580	5437	2863	2749	36	36	0	0	37	37
South Korea	3	4816	4872	7657	7976	2841	3104	697	697	10943	10943	10247	10247
Bahrain	6	2594	2654	5399	5388	2805	2746	3	3	27	27	23	23
Oman	6	4370	4603	5041	5822	2794	2770	0	0	333	333	333	333
Australia	93	1651	737	4193	1279	2779	373	3181	1687	4805	1748	4261	1852
Singapore	18	375	116	2940	1608	2565	1244	78	0	5060	16	4982	16
Malaysia	9	746	429	3305	1091	2558	673	0	0	1220	1205	1220	1205

Table 4.6: CDN top 10 countries sorted by largest CBG error distances (km) difference (EDD)

From Figure 4.6 we see that IPv4 CBG performs slightly better than IPv6 CBG in estimating the distance to the target. For approximately 40% of the targets, CBG IPv4 outperforms MaxMind, this occurs around 1000km. CBG IPv6 outperforms MaxMind.

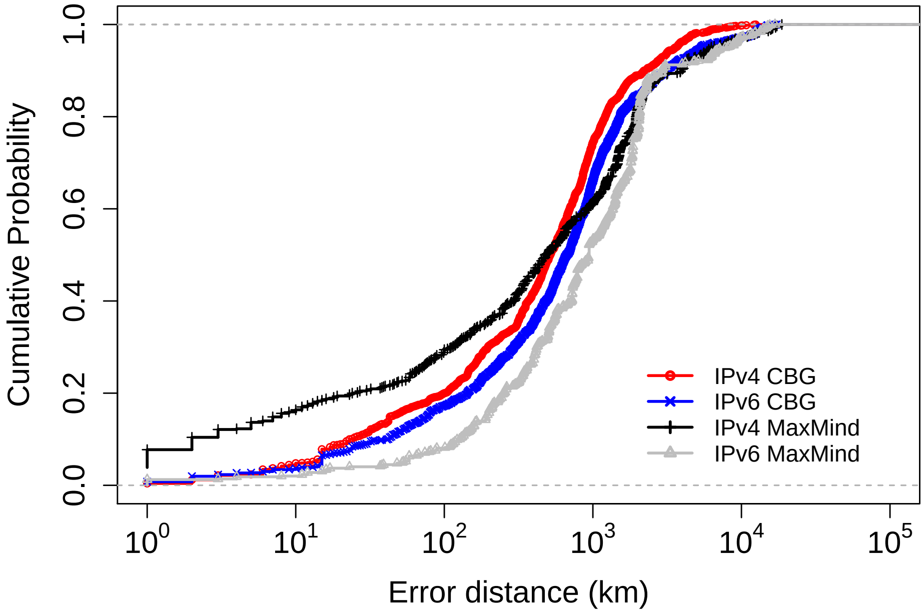

Figure 4.6: CDF error distance for CDN CBG and MaxMind

Figure 4.7 shows the CDF of the EDD for CBG and MaxMind. For roughly 60% of the targets, CBG outperforms MaxMind, this occurs around 100km.

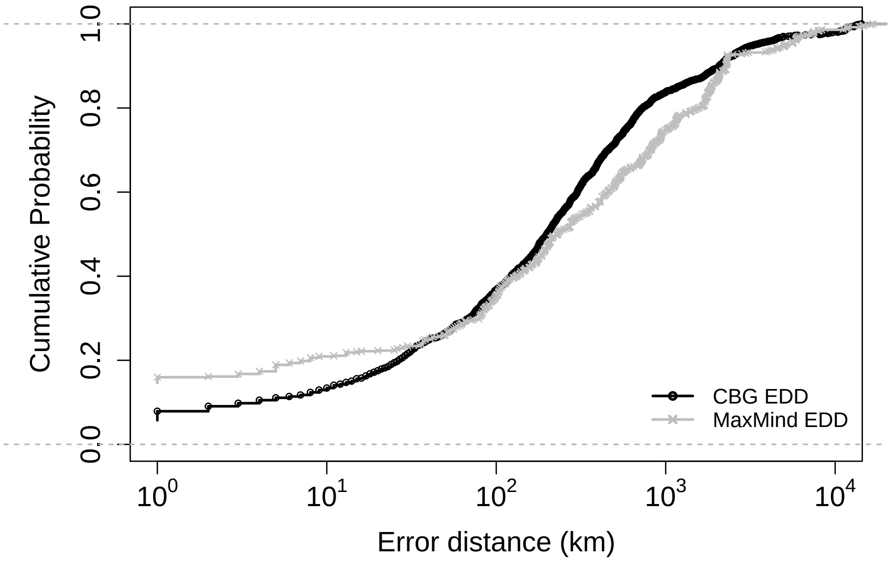

Figure 4.7: CDF of CDN error distance difference (EDD) for CBG and MaxMind

Figure 4.8 shows the confidence regions for the CDN dataset. Similar to what we found in the UNI dataset, the IPv4 areas are consistently smaller than the IPv6 regions, thus giving us a higher confidence of target host location estimate. Results show that for about 50% of the areas for IPv4 and IPv6 are about the size of the U.S. state of Texas.

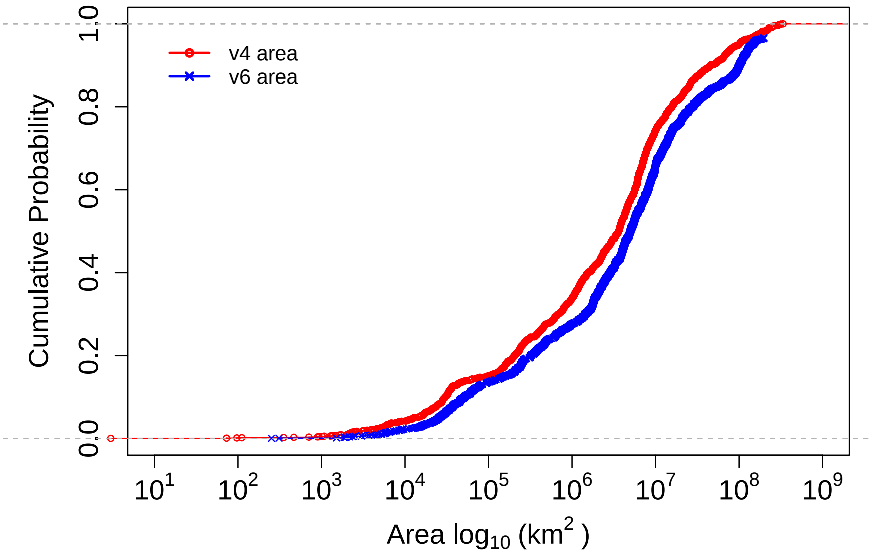

Figure 4.8: CDF Confidence region for CDN v4-v6 pairs

4.4.2 CDN AS-level Path

For the CDN dataset, 54,682 IPv4 and 58,026 IPv6 traceroutes were collected from our landmarks during our ping probing sessions. Like our UNI dataset, we discard duplicate traceroutes between a landmark and host, resulting in our final set of traceroutes to perform our analysis. Of those received, 3,472 IPv4 and 12,592 IPv6 traceroutes did not reach the target hosts. We were able to match 45,409 traceroutes between IPv4 and IPv6 for direct comparison of AS-level edit distance and PL. This is approximately an 84% match rate to IPv4 traceroutes.

Table 4.7 shows the number of traces per country and region. We can see that the average edit distances and path lengths are lowest for RIPE and ARIN, both below the overall edit distance average. These are also regions where the CBG performed the best as we found in the Section 4.4.1. We find that the Asia region average edit distances require the most edits to match IPv4 AS-level paths to their IPv6 counterpart. This may explain CBG's subpar performance although we have five landmarks in the APNIC region. Additionally, the overall IPv6 PL is shorter than the overall IPv4 PL. As mentioned in Section 2.2.3, recent work has shown that the IPv6 PL has been decreasing over time, and in some cases has become shorter than IPv4 PL. Although this may be the case, shorter PL does not appear to affect CBG performance as much as edit distances.

Location	No. of Traces	Edit Distance Average	Edit Distance Median	IPv4 PL Average	IPv6 PL Median
AfriNIC	667	2.583	3	4.154	4.495
APNIC	4237	2.72	3	4.207	4.156
RIPE NCC	15436	2.005	2	3.681	3.747
LACNIC	2260	2.686	3	4.146	4.103
ARIN	19128	2.068	2	3.749	3.691
Oceania	3681	2.517	2	4.342	4.444
Total	45409	2.182	2	3.842	3.84

Table 4.7: CDN edit distances and path lengths

4.5 One-to-One (ONE2ONE) Pairs Dataset

Our last dataset comes from [13] and includes IPv4 and IPv6 addresses collected in January 2013. The locations of the ONE2ONE pairs are unknown. However, we use the MaxMind inferred country and region for a point of reference for evaluation. We initially had 3,417 IPv4-v6 address pairs. After executing ping probes to all of these address pairs from our landmarks, we received 1,691 matching active response pairs. Since true locations are unknown, we compare the error distance between the geographic coordinate centroids for CBG IPv4 and IPv6. This is also done for MaxMind estimated locations. We note that 34 IPv6 addresses were not found in the MaxMind database.

4.5.1 ONE2ONE Geolocation

We see from Table 4.8 that the EDD distance and average is smallest for RIPE NCC & ARIN regions. The CBG EDD average and median is largest in the LACNIC region. As previously noted in the UNI dataset, the inferred MaxMind location for both IPv4 and IPv6 may be assigned to the same geographic coordinates, driving the MaxMind EDD much lower. This appears to be the case for APNIC and LACNIC regions since the EDD median is surprisingly low for this dataset, while we found in the CDN dataset indications that MaxMind IPv4 and IPv6 location estimates are very large from each other. From Table 4.9, we see that LACNIC has a median edit distance of four which is an indicator of this occurring. Also, the highest average edit distance is from the LACNIC region and it has the only median edit distance of four among all the regions. This supports the finding that the greater the edit distances, the greater the error distance.

		CBG		***MaxMind***	
Location (MaxMind Inferred)	No. of Hosts	EDD Average	EDD Median	EDD Average	EDD Median
AfriNIC	8	3645	3828	1477	431
APNIC	204	3899	1209	150	5
RIPE NCC	896	319	182	365	82
LACNIC	19	7411	8184	85	0
ARIN	509	1070	97	1464	1843
Oceania	55	5989	2785	814	299
Total	1691	1257	218	686	127

Table 4.8: ONE2ONE regional error distance differences (km)

Figure 4.9 further shows that the MaxMind EDD has smaller error distances overall compared to CBG.

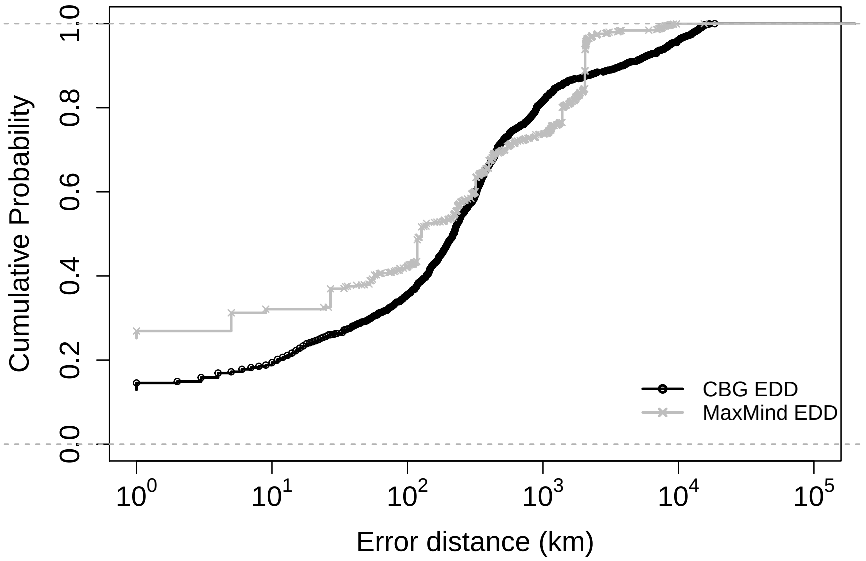

Figure 4.9: CDF error distance difference (EDD) ONE2ONE v4-v6 pairs

As shown in Figure 4.10, the IPv6 area size is comparable to IPv4 area. Areas larger than 10^6 km^2, a little larger than the area of Texas, is where IPv6 confidence regions begins to diminish compared to IPv4.

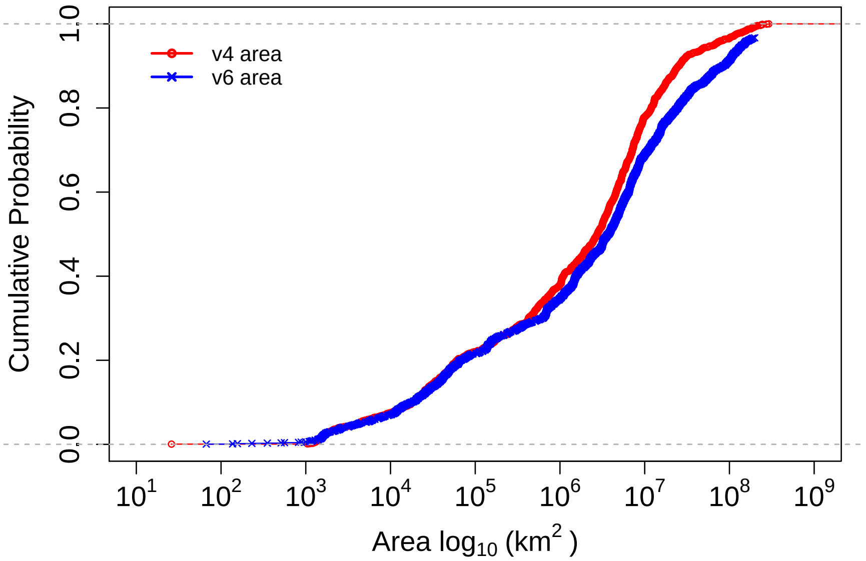

Figure 4.10: CDF area regions for ONE2ONE

4.5.2 ONE2ONE AS-level Paths

We received 47,821 IPv4 and 50,682 IPv6 traceroutes from our landmarks. Of those received, 3,055 IPv4 and 11,170 IPv6 traceroutes did not reach the target hosts. We were able to match 39,489 traceroutes between IPv4 and IPv6 for direct comparison of AS-level edit distance and PL. This is roughly 82% match rate to total IPv4 traceroutes.

Here again we find that LACNIC and APNIC region average edit distances are the highest among its peers. Looking back at Table 4.8 we see that LACNIC has the largest EDD median distance, but is followed by Oceania and then APNIC for largest inferred location median difference distance.

Location (MaxMind Inferred)	No. of Traces	Edit Distance Average	Edit Distance Median	IPv4 PL Average	IPv6 PL Average
AfriNIC	198	2.48	3	5.323	5.283
APNIC	4976	2.907	3	4.364	4.015
RIPE NCC	21678	2.443	2	4.047	3.935
LACNIC	335	3.678	4	4.988	4.191
None	26	1.769	1	3.846	3.538
ARIN	11014	2.176	2	3.756	3.351
Oceania	1262	2.61	2	4.677	4.595
Total	39441	2.443	2	4.04	3.812

Table 4.9: ONE2ONE regional edit distances with path length

CHAPTER 5: Conclusion

This thesis sought to investigate whether using latency-based measurements was a viable "first-step" coarse-grain geolocation technique for IPv6 hosts. Using the CBG technique from Gueye et al. we created bestline models using multilateration of known landmarks to capture the network delay patterns of the IPv4 and IPv6 Internet. Using great circle constraints from our landmarks, we geolocated target hosts to a constrained area to produce an inferred target location. We performed CBG geolocation for IPv4 and IPv6 pairs to compare performance as well as provide a general comparison against a commercial IP geolocation database. We further provided some insight into performance variations by computing edit distances and path lengths from traceroutes between our set of landmarks and each target host.

We found that among all datasets, overall, IPv6 CBG geolocated targets with error distances and area regions larger than IPv4 CBG, but some regions had much greater error distances than others. The worst case was the Oceania region where CBG IPv6 median error distance was doubled as compared to CBG IPv4 median error distance. The best case was the RIPE region where CBG IPv6 median error distance outperformed CBG IPv4 median error distance by roughly 8%. The major differences in the two regions, Oceania and RIPE NCC, are the AS-level differences found between IPv4 and IPv6 in those regions, and the landmark density. Consistent with previous IP geolocation studies leveraging landmarks, we also find that, typically, the higher the density of landmarks in a region of the target host, the higher the geolocation accuracy [9, 24–26, 29]. We also found that targets located in the Oceania region had among the largest EDD between IPv4 and IPv6 even though four landmarks were present in that region. For UNI and CDN datasets with known host locations, hosts in Oceania had an IPv4 error distance average of at least 1,800km.

Using AS-level paths between landmarks and hosts, we used edit distances and PLs to provide insight into our geolocation performance. We found in each dataset that edit distances seemed to have a direct relationship with CBG accuracy. The greater the edit distances, the greater the error distance to the target hosts. We calculated the PCC for edit distance to median error distance for UNI and found IPv4 had 0.997 correlation and IPv6 had 0.994 correlation. For the CDN dataset, the PCC for edit distance to median error distance of IPv4 was 0.584 correlation and IPv6 was 0.860 correlation. PL did not show a strong relationship with CBG accuracy.

5.1 Future Work

To the best of our knowledge, this thesis is the first publicly available latency-based IPv6 geolocation study using a modest probing infrastructure and non-trivial target dataset. Below are items that could be studied to further this initial work:

Increase the number of landmarks in specific regions. Our study did not have any vantage points or landmarks located in the AfriNIC or LACNIC regions. Increasing the number of landmarks in both regions could boost CBG geolocation performance. Our results show that areas where we did have landmarks generally resulted in higher CBG geolocation performance; however, this was not always the case. The APNIC and Oceania regions each had four landmarks, yet CBG geolocation EDD for the UNI and CDN datasets were significantly larger compared to other regions with landmarks. It would be valuable to increase the number of landmarks in the APNIC and Oceania regions, and observe the change to CBG geolocation. Adding landmarks in AfriNIC and LACNIC would also show if CBG geolocation would perform as well as the RIPE NCC and ARIN regions did on our data.

Selectively choose landmarks. A study conducted by [29] showed that with many vantage points, vantage point proximity to the target host is the most important factor affecting accuracy. A process to select only the *best* landmarks to geolocate host could be developed. We suggest two features as a starting point for review: RTT and correlation coefficients.

First, in selecting the best set of landmarks among a pool of landmarks to geolocate a host, we suggest using landmarks based on a minimum RTT threshold. In other words, after sorting RTTs measured among a set of landmarks to a target host from lowest to highest RTT, determine a cut off threshold of minimum RTTs to use in geolocating the target host. The cut off RTT threshold would vary depending on which landmarks further minimized the error distance between CBG target host estimation to true host location. This suggestion is based on the observation that short RTTs between a landmark and target host typically show landmarks that are geographically closer to target. If a RTT threshold were set for each target host, only landmarks that were close to the target host would be used to geolocate the host. This falls in line with the study completed by [26] that geolocation rarely works better than the distance to the nearest landmark. If a much larger set of landmarks were accessible, using this RTT threshold strategy could potentially increase geolocation accuracy.

The second feature where we suggest further study is how correlation coefficients of landmark bestline models affect geolocation performance. We found Ark landmark bestline models with negative Pearson correlation coefficients had poor RTT to geographical distance relationship. In our study, we did not discard any Ark landmarks when geolocating hosts. However, learning what effects, if any, if a landmark with poor correlation affects geolocation, would be interesting. If this feature does prove beneficial, selecting a set of landmarks among a larger set could further increase CBG geolocation accuracy.

Investigate delay within AS. From Section 4 we found that AS edit distances between IPv4 and IPv6 paths have a strong relationship on CBG geolocation accuracy. We also found in some regions where IPv4 or IPv6 path lengths were shorter compared to other regions, yet experienced larger error distances. It would be interesting to conduct more analysis on delay within certain ASs. This could provide CBG geolocation an ability to compensate, or adjust calculations, when ping probes are measured through ASs that have longer routing times.

Fine-grain geolocation. The study conducted in [9] developed a three-tier model for fine-grain geolocation, decreasing error distances to as low as 690 meters from estimated location to true location. The researchers' results were based only on IPv4 hosts and a large pool of over 160 ping and traceroute servers. Additionally, their study collected web-based landmarks that contributed to their geolocation. If able to acquire a comparable pool of IPv6 capable ping and traceroute servers along with IPv6 capable web-based landmarks, it would be interesting to see the ability to conduct fine-grain geolocation in the IPv6 space. Moreover, if able to leverage IPv4 web-based landmarks, it would be valuable to understand how this added information could support IPv6 geolocation in a non-dense IPv6 environment.

THIS PAGE INTENTIONALLY LEFT BLANK

REFERENCES

[1] Internet Corporation for Assigned Names and Numbers, “Available pool of unallocated IPv4 Internet addresses now completely emptied,” 2011. [Online]. Available: www.icann.org/en/news/releases/release-03feb11-en.pdf

[2] G. Huston, “IPv4 address report,” 2013. [Online]. Available: http://www.potaroo.net/tools/ipv4/index.html

[3] P. Srisuresh and K. Egevang, “Traditional IP Network Address Translator (Traditional NAT),” RFC 3022, Jan. 2001.

[4] T. Hain, “Architectural implications of NAT,” RFC 2993, Nov. 2000.

[5] Google, “Google IPv6 adoption,” 2013. [Online]. Available: http://www.google.com/ipv6/statistics.html#tab=ipv6-adoption

[6] RIPE-NCC, “IPv6 enabled networks,” 2012. [Online]. Available: http://v6asns.ripe.net/v/6

[7] S. Deering and R. Hinden, “Internet Protocol, version 6 (IPv6) specification,” RFC 2460 (Draft Standard), Dec. 1998.

[8] The Federal CIO Council Strategy and Planning Committee, “Planning guide/roadmap toward IPv6 adoption within the U.S. government,” Washington D.C., July 2012.

[9] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, and C. Huang, “Towards street-level client-independent IP geolocation,” in *Proc. of the 8th USENIX Conference on Networked Systems Design and Implementation*, 2011, pp. 27–36.

[10] K. Claffy, “Tracking IPv6 evolution: Data we have and data we need,” *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 3, pp. 43–48, July, 2011.

[11] A. Dhamdhere, M. Luckie, B. Huffaker, K. Claffy, A. Elmokashfi, and E. Aben, “Measuring the deployment of IPv6: topology, routing and performance,” in *Proc. of the 2012 ACM Internet Measurement Conference*, 2012, pp. 537–550.

[12] Ø. E. Thorvaldsen, "Geographical Location of Internet Hosts using a Multi-Agent System," Ph.D. dissertation, Dept. Comp. and Info. Science, Norwegian Univ., Trondheim, Norway, 2006.

[13] A. Berger, N. Weaver, R. Beverly, and L. Campbell, "Internet nameserver IPv4 and IPv6 address relationships," in *Proc. of the ACM SIGCOMM Internet Measurement Conference (IMC)*, 2013 [Online]. Available: http://rbeverly.net/research/papers/dnsv4v6-imc13.html

[14] I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye, "IP geolocation databases: unreliable?" *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 2, pp. 53–56, Apr. 2011 [Online]. Available: http://doi.acm.org/10.1145/1971162.1971171

[15] MaxMind, Inc, "MaxMind," 2013. [Online]. Available: http://www.maxmind.com/en/home

[16] Y. Shavitt and N. Zilberman, "A geolocation databases study," *Selected Areas in Communications, IEEE Journal on Selected Areas in Communications*, vol. 29, no. 10, pp. 2044–2056, 2011.

[17] R. Koch, M. Golling, and G. D. Rodosek, "Advanced Geolocation of IP Addresses," *International Journal of Electrical, Electronic Science and Engineering*, vol. 7, no. 2, pp. 1–10, 2013. [Online]. Available: http://waset.org/Publications?p=80

[18] B. Huffaker, M. Fomenkov, and K. Claffy, *Geocompare: A comparison of public and commercial geolocation databases*, Cooperative Association for Internet Data Analysis (CAIDA), Tech. Rep., May 2011.

[19] L. Daigle, "WHOIS Protocol Specifications," RFC 3912, Sept. 2004.

[20] P. Mockapetris, "Domain names - concepts and facilities," RFC 1034, Nov. 1987.

[21] R. Landa Gamiochipi, J. Araujo, R. Clegg, E. Mykoniati, D. Griffin, and M. Rio, "The Large-Scale Geography of Internet Round Trip Times," in *Proc. of IFIP Networking*, 2013.

[22] G. Kessler and S. Shepard, "A primer on Internet and TCP/IP tools and utilities," RFC 2151, June 1997.

[23] R. Percacci and A. Vespignani, "Scale-free behavior of the Internet global performance," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 32, no. 4, pp. 411–414, 2003.

[24] V. N. Padmanabhan and L. Subramanian, “An investigation of geographic mapping techniques for Internet hosts,” *ACM SIGCOMM Computer Communication Review*, vol. 31, no. 4, pp. 173–185, 2001.

[25] B. Gueye, M. Crovella, A. Ziviani, and S. Fdida, “Constraint-based geolocation of Internet hosts,” *IEEE/ACM Transactions on Networking*, vol. 14, no. 6, pp. 1219–1232, 2006.

[26] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, “Towards IP geolocation using delay and topology measurements,” in *Proc. of the 6th ACM SIGCOMM Internet Measurement Conference*, 2006, pp. 71–84.

[27] G. Huston, “IPv6: IPv6 / IPv4 Comparative Statistics,” 2014. [Online]. Available: http://bgp.potaroo.net/v6/v6rpt.html

[28] J. Taylor, J. Devlin, and K. Curran, “Bringing location to IP addresses with IP geolocation,” *Journal of Emerging Technologies in Web Intelligence*, vol. 4, no. 3, pp. 273–277, 2012.

[29] Z. Hu, J. Heidemann, and Y. Pradkin, “Towards geolocation of millions of IP addresses,” in *Proc. of the 2012 ACM Internet Measurement Conference*. 2012, pp. 123–130.

[30] P. Misra and P. Enge, “Special Issue on Global Positioning System,” in *Proc. of the IEEE*, vol. 87, no. 1, pp. 3–15, 1999.

[31] The Number Resource Organization, “Regional Internet Registries,” 2014. [Online]. Available: http://www.nro.net/about-the-nro/regional-internet-registries

[32] Cooperative Association for Internet Data Analysis, “Archipelago Measurement Infrastructure,” 2013. [Online]. Available: http://www.caida.org/projects/ark

[33] R. G. Chamberlain and W. H. Duquette, “Some algorithms for polygons on a sphere,” presented at the Association of American Geographers Annual Meeting, San Francisco, CA, 2007.

[34] University of Oregon, “Route Views,” Advanced Network Technology Center, 2014. [Online]. Available: http://routeviews.org

[35] M. Luckie et al., “A second look at detecting third-party addresses in traceroute traces with the ip timestamp option,” in *Passive and Active Measurement*. New York: Springer, 2014, pp. 46–55.

[36] K. Pearson, “Note on regression and inheritance in the case of two parents,” in *Proc. Royal Society of London*, 1895, vol. 58, no. 347-352, pp. 240–242.

[37] B. Render, R. M. Stair, and M. Hanna, *Quantitative Analysis for Management*, 11th ed. New York, NY: Pearson, 2012.

[38] K. Cho, M. Luckie, and B. Huffaker, “Identifying ipv6 network problems in the dual-stack world,” in *Proceedings of the ACM SIGCOMM workshop on Network troubleshooting: research, theory and operations practice meet malfunctioning reality*. ACM, 2004, pp. 283–288.

[39] AllesEDV, “IP Reachability,” 2013. [Online]. Available: www.allesedv.at/IPv6/tld/.edu

[40] latlon.net, “Latitude Longitude Finder,” 2013. [Online]. Available: http://www.latlong.net/convert-address-to-lat-long.html

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia
2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California